An Exponential Tail Bound for the Deleted Estimate

Karim Abou–Moustafa[*] and **Csaba Szepesvári**[†]
Dept. of Computing Sciecne, University of Alberta
Edmonton, Alberta T6G 2E8, Canada
aboumous@ualberta.ca, szepesva@ualberta.ca

Abstract

There is an accumulating evidence in the literature that *stability of learning algorithms* is a key characteristic that permits a learning algorithm to generalize. Despite various insightful results in this direction, there seems to be an overlooked dichotomy in the type of stability-based generalization bounds we have in the literature. On one hand, the literature seems to suggest that exponential generalization bounds for the estimated risk, which are optimal, can be *only* obtained through *stringent*, *distribution independent* and *computationally intractable* notions of stability such as *uniform stability*. On the other hand, it seems that *weaker* notions of stability such as hypothesis stability, although it is *distribution dependent* and more *amenable* to computation, can *only* yield polynomial generalization bounds for the estimated risk, which are suboptimal. In this paper, we address the gap between these two regimes of results. In particular, the main question we address here is *whether it is possible to derive exponential generalization bounds for the estimated risk using a notion of stability that is computationally tractable and distribution dependent, but weaker than uniform stability*. Using recent advances in concentration inequalities, and using a notion of stability that is weaker than uniform stability but distribution dependent and amenable to computation, we derive an exponential tail bound for the concentration of the estimated risk of a hypothesis returned by a *general* learning rule, where the estimated risk is expressed in terms of the deleted estimate. Interestingly, we note that our final bound has similarities to previous exponential generalization bounds for the deleted estimate, in particular, the result of Bousquet and Elisseeff (2002) for the regression case.

1 Introduction

There is an accumulating evidence in the literature that stability of learning algorithm is a key characteristic that permits a learning algorithm to generalize. The earliest result in this regard is due to Devroye and Wagner (1979a) and Devroye and Wagner (1979b) where they derive distribution free *polynomial* generalization bounds for the concentration of the *leave-one-out* estimate, or the *deleted* estimate, for the expected error of some nonparametric learning rules. Although the notion of stability was not explicitly mentioned in their work, the polynomial bounds of Devroye and Wagner (1979b) for instance relied on the so called *hypothesis stability*; a name that is due to Kearns and Ron (1999). Various results for different estimates followed the works of Devroye and Wagner (1979b). Lugosi and Pawlak (1994) extended the work of Devroye and Wagner (1979b) to smooth estimates of the error developed in terms of *a posteriori* distribution for the deleted estimate. Holden (1996) derived *sanity-check bounds* for the deleted estimate and the *k folds cross–validation* (KFCV) estimate using hypothesis stability for few algorithms in the realizable setting. Kearns and Ron (1999), using the notion of *error stability*, give sanity-check bounds for the deleted estimate but for more general classes of learning rules (in the unrealizable or agnostic setting). More recently, Kale, Kumar, and Vassilvitskii (2011) show that, using a weak notion of stability known as *mean-square* stability, the averaging taking place in the KFCV estimation procedure can reduce the variance of the generalization error; i.e. the averaging in the KFCV estimation procedure can improve the concentration of the estimated error around the expected error of the hypothesis returned by the learning rule.

For general learning rules and for *regularized empirical risk minimization* learning rules, Bousquet and Elisseeff (2002) using the notion of *uniform stability*, extended the work of Lugosi and Pawlak (1994) and derived *exponential* generalization bounds for the resubstitution estimate and the deleted estimate. Further generalization results based on uniform stability (or one of its variants) were obtained in the works of Kutin and Niyogi; Rakhlin, Mukherjee, and Poggio; Mukherjee et al.; Shalev-Shwartz et al. (2002; 2005; 2006; 2010), to name but a few. These results were extended in various directions such as deriving new results for randomized learning algorithms (Elisseeff, Evgeniou, and Pontil 2005), transfer and meta learning (Maurer 2005), adaptive data analysis (Bassily et al. 2016), stochastic gradient descent (Hardt, Recht, and Singer 2016), structured prediction (London, Huang, and Getoor 2016), multi-task learning (Zhang 2015), ranking algorithms (Agarwal and Niyogi 2009), as well as in understanding the trade-off between sparsity and stability (Xu, Caramanis, and Mannor 2012).

Despite these recent advances, and excluding sanity-check bounds, there seems to be an overlooked dichotomy in the type of stability-based generalization results. In particu-

[*] Currently with SAS Inst. Inc., Cary, North Carolina, USA

[†] Currently with Google DeepMind, London, UK

lar, the results on stability and generalization can be grouped into two regimes:

1. *Polynomial* generalization bounds, which are *sub-optimal* and based on hypothesis stability for instance.
2. *Exponential* generalization bounds, which are *optimal* and based on uniform stability (and its variants).

Comparing *uniform* stability to other notions of stability in the literature, uniform stability is the strongest notion of stability in the sense that it implies all other notions of stability such as hypothesis stability, error stability, and mean-square stability (Bousquet and Elisseeff 2002). A learning rule is *uniformly stable* if the change in the prediction loss is small, no matter how the input to the learning rule is selected, no matter what value is used as a test example, and no matter which example is removed (or replaced) in the input.

Despite the strength of uniform stability, it is unpleasantly restrictive. First, unlike other notions of stability (e.g. L_2 and L_1 stability), uniform stability is a stringent notion of stability that is insensitive to the data-generating distribution. This is problematic since it removes the possibility of studying large classes of learning rules, or even classes of problems. One particularly striking example is binary classification with the zero-one loss. For this problem, as it was already noted by (Bousquet and Elisseeff 2002), *no trivial algorithm* can be uniformly β-stable with $\beta < 1$. Another example when uniform stability fails is regression with unbounded losses and response variables. Second, as noted earlier, uniform stability is distribution-free and is thus unsuitable for studying finer details of learning algorithms. Computation is another aspect that distinguishes uniform stability from other notions of stability. While hypothesis, error, and mean-square stability can be estimated using a finite sample, uniform stability is computationally intractable which is problematic if it is desired to obtain *empirical high probability generalization bounds* for the expected risk in the spirit of empirical Bernstein bounds for instance (Audibert, Munos, and Szepesvári 2007; Mnih, Szepesvári, and Audibert 2008).

In this research, we are particularly motivated by these previous observations. That is, on the one hand, the literature seems to suggest that exponential generalization bounds for the estimated risk, which are optimal, can be *only* obtained through *stringent*, *distribution independent*, and *computationally intractable* notions of stability such as uniform stability (and its variants). On the other hand, it seems that *weaker* notions of stability such as hypothesis and mean-square stability, although they are *distribution dependent* and more *amenable* to computation, can *only* yield polynomial generalization bounds for the estimated risk, which are sub-optimal.

The chief purpose of this paper is to address the gap between these two regimes of results. In particular, the main question we address here is *whether it is possible to derive exponential generalization bounds for the estimated risk using a notion of stability that is computationally tractable, distribution dependent, but weaker than uniform stability*. In this work, we show that using recent advances in exponential concentration inequalities, and using a notion of stability that is distribution dependent, amenable to computation, but weaker than uniform stability, we derive in Theorem 4 an exponential tail bound for the concentration of the estimated risk of a hypothesis returned by a *general* learning rule, where the estimated risk is developed in terms of the deleted estimate. Interestingly, we note that our final bound has similarities to previous exponential generalization bounds for the deleted estimate, in particular the result of Bousquet and Elisseeff (2002) for the regression case.

Two main ingredients that allowed us to bridge the gap between these two regimes of results; (*i*) recent advances in exponential concentration inequalities, in particular the exponential Efron-Stein inequality due to Boucheron, Lugosi, and Massart (2003) and Boucheron, Lugosi, and Massart (2013); and (*ii*) the elegant and smart notion of L_q stability due to Celisse and Guedj (2016) which is distribution dependent, computationally tractable, but weaker than uniform stability, and generalizes hypothesis stability and mean-square stability to higher order moments.

2 Setup and Notations

We consider learning in Vapnik's framework for risk minimization with bounded losses (Vapnik 1995): A learning problem is specified by the triplet $(\mathcal{H}, \mathcal{X}, \ell)$, where $\mathcal{H}, \mathcal{X}$ are sets and $\ell : \mathcal{H} \times \mathcal{X} \to [0, 1]$. The set $\mathcal{H}$ is called the *hypothesis space*, $\mathcal{X}$ is called the *instance space*, and ℓ is called the *loss function*. The loss $\ell(h, x)$ indicates how well a hypothesis h explains (or fits) an instance $x \in \mathcal{X}$. The learning problem is defined as follows. A learner $\mathtt{A}$ sees a sample in the form of a sequence $\mathcal{S}_n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ where $(X_i)_i$ is sampled in an independent and identically distributed (*i.i.d*) fashion from some unknown distribution $\mathscr{P}$ and returns a hypothesis $\widehat{h}_n = \mathtt{A}(\mathcal{S}_n) \in \mathcal{H}$ based solely on $X_1, \ldots, X_n$.[1] The goal of the learner is to pick hypotheses with a small *risk* (defined shortly).

We assume that a learner is able to process samples (or sequences) of different cardinality. Hence, a learner will be identified with a map $\mathtt{A} : \cup_n \mathcal{X}^n \to \mathcal{H}$. We only consider deterministic learning rules in this work; the extension to randomizing learning rules is left for future work.

Given a distribution $\mathscr{P}$ on $\mathcal{X}$, the risk of a *fixed hypothesis* $h \in \mathcal{H}$ is defined to be $R(h, \mathscr{P}) = \mathbb{E}\left[\ell\left(h, X\right)\right]$, where $X \sim \mathscr{P}$. Since $\mathcal{S}_n$ is a random quantity, so are $\mathtt{A}(\mathcal{S}_n)$ and $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$, the latter of which can be also written as $\mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) |\mathcal{S}_n]$, where $X \sim \mathscr{P}$ is independent of $\mathcal{S}_n$. Ideal learners keep the risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$ of the hypothesis returned by $\mathtt{A}$ "small" for a wide range of distributions $\mathscr{P}$.

q-Norm of Random Variables: In the sequel, we will heavily rely on the q-norm for random variables (RVs). For a real RV X, and for $1 \le q \le +\infty$, the q-norm of X is defined as: $\left\|X\right\|_q \doteq \left(\mathbb{E}\left[|X|^q\right]\right)^{1/q}$, and $\left\|X\right\|_\infty$ is the essential supremum of $|X|$. Note that for $1 \le q \le p \le +\infty$, these norms satisfy $\left\|\cdot\right\|_q \le \left\|\cdot\right\|_p$.

[1] The set $\mathcal{X}$ is thus measurable. In general, to minimize clutter, all functions are assumed to be measurable as needed.

2.1 Risk Estimators

The generalization bounds on the risk usually center on some point-estimate of the random risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$. Many estimators are based on calculating the sample mean of losses in one form or another. For any fixed hypothesis $h \in \mathcal{H}$ and dataset $\mathcal{S}_n$, the sample mean of losses of h against $\mathcal{S}_n$, also known as the *empirical risk* of h on $\mathcal{S}_n$, is given by

$$\widehat{R}(h, \mathcal{S}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(h, X_i\right). \tag{1}$$

Plugging $\mathtt{A}(\mathcal{S}_n)$ into $\widehat{R}(\cdot, \mathcal{S}_n)$ we get the *resubstitution (RES) estimate*, or the training error (Devroye and Wagner 1979b): $\widehat{R}_{\text{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) = \widehat{R}\left(\mathtt{A}\left(\mathcal{S}_n\right), \mathcal{S}_n\right)$. The resubstitution estimate is often overly "optimistic", i.e., it underestimates the actual risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$. The *deleted (DEL) estimate* defined as

$$\widehat{R}_{\text{DEL}}\left(\mathtt{A}, \mathcal{S}_n\right) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right), \tag{2}$$

is a common alternative to the resubstitution estimate that aims to correct for this optimism. Here, $\mathcal{S}_n^{-i} = (X_1, \ldots, X_{i\text{-}1}, X_{i+1}, \ldots, X_n)$, i.e., it is the sequence $\mathcal{S}_n$ with example X_i removed. Since $\mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)] = R_{n-1}(\mathtt{A}, \mathscr{P})$, then $\mathbb{E}[\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)] = R_{n-1}(\mathtt{A}, \mathscr{P})$. When the latter is close to $R_n(\mathtt{A}, \mathscr{P})$, i.e., $\mathtt{A}$ is "stable", the deleted estimate may be a good alternative to the resubstitution estimate (Devroye, Györfi, and Lugosi 1996). However, due to the potentially strong correlations between elements of $(\ell(\mathtt{A}(\mathcal{S}_n^{-i}), X_i))_i$, the variance of the deleted estimate *may be* significantly higher than that of the resubstitution estimate due to the overly redundant information content between $\ell(\mathtt{A}(\mathcal{S}_n^{-i}), X_i)$ and $\ell(\mathtt{A}(\mathcal{S}_n^{-j}), X_j)$ for $i \neq j$. The main goal of this work is to develop a high probability upper bound on the absolute deviation $|\widehat{R}_{\text{DEL}}\left(\mathtt{A}, \mathcal{S}_n\right) - R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)|$ in terms of the "stability" of $\mathtt{A}$, which is defined next.

3 Notions of Stability for Learning Rules

In the following, and due to space limitations, we go briefly over two notions of algorithmic stability that will be needed in our context; namely the uniform stability due to Bousquet and Elisseeff (2002), and the L_q stability due to Celisse and Guedj (2016).

Definition 1 (Uniform Stability)**.** *Algorithm* $\mathtt{A}$ *has uniform stability* β_u *w.r.t the loss function* ℓ *if the following holds:* $\forall\, \mathcal{S}_n \in \mathcal{X}^n$, $\forall i \in \{1, \ldots, n\}$,

$$\|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\|_\infty \;\leq\; \beta_u\,,$$

where $X \sim \mathscr{P}$ *is independent of* $\mathcal{S}_n$*.*

Definition 2 (L_q Stability Coefficient)**.** *Let* $\mathcal{S}_n$ *be a sequence of* n *i.i.d random variables (RVs) drawn from* $\mathcal{X}$ *according to* $\mathscr{P}$*. Let* $\mathtt{A}$ *be a deterministic learning rule, and* ℓ *be a loss function as defined in Section 2. For* $q \geq 1$*, the* L_q stability *coefficient of* $\mathtt{A}$ *w.r.t* ℓ*,* $\mathscr{P}$*, and* n *is denoted by* $\beta_q(\mathtt{A}, \ell, \mathscr{P}, n)$ *and is defined as*

$$\beta_q^q(\mathtt{A}, \ell, \mathscr{P}, n) \;=\; \mathbb{E}\left[\left|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X\right)\right|^q\right]\,,$$

where $X \sim \mathscr{P}$ *is independent of* $\mathcal{S}_n$*.*

In our definition of L_q stability coefficients, we simply assume that it is always the first example that is removed. This is because the examples in $\mathcal{S}_n$ are *i.i.d*, and hence the joint distribution of $(\mathtt{A}(\mathcal{S}_n), \mathtt{A}(\mathcal{S}_n^{-1}), X)$ does not depend on which example is removed from $\mathcal{S}_n$. Note that quite a few previous works restrict notions of algorithmic stability to learning rules that are permutation invariant, or *"symmetric"*; i.e. learning rules that yield identical output under different permutations of the examples presented to them (Rogers and Wagner 1978; Devroye and Wagner 1979b; Bousquet and Elisseeff 2002; Shalev-Shwartz et al. 2010). For the same reason of why it does not matter which examples are removed, it does not matter whether the learning rule is symmetric or not.

Since often $\mathtt{A}$, ℓ, $\mathscr{P}$ are fixed, we will drop them from the notation and will just use $\beta_q^q(n)$.[2] Note that when $q = 1$ and $q = 2$, L_q stability reduces to hypothesis stability (Devroye and Wagner 1979b) and MS stability (Kale, Kumar, and Vassilvitskii 2011), respectively, which were introduced earlier in the literature. Note also that uniform stability implies L_q stability for every $q \geq 1$. The L_q stability coefficient quantifies the variation of the loss of $\mathtt{A}$ induced by removing one example from the training set. This is known as a *removal type* notion of stability and is in accordance with the previous notions of stability introduced earlier. The difference between L_q stability and earlier notions of stability is that L_q stability is in terms of the higher order moments of the RV $|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X\right)|$. The reason we care about higher moments is because we are interested in controlling the tail behavior of the deleted estimate. As will be shown, the tail behavior of the deleted estimate is also dependent on the tail behavior of RVs characterizing stability. As is well-known, knowledge of the higher moments of a RV is equivalent to knowledge of the tail behavior of the RV. As such, controlling the higher order moments provides more information on the distribution of this RV than simply considering first order (L_1) and second order (L_2) moments.

From the definition of q–Norm of a RV in Section 2, observe that the stability coefficient $\beta_q(n) \doteq \beta_q(\mathtt{A}, \ell, \mathscr{P}, n)$ is in fact a q–norm for the RV: $\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X\right)$. From the properties of q-norms, we have that for $1 \leq q \leq p \leq +\infty$, it holds that $\beta_q(n) \leq \beta_p(n)$; i.e. $\beta_q(n)$ is an increasing function of q. Furthermore, we expect that $\beta_q(n)$ has the following trend as a function of n

Assumption 1. *For a fixed* $q > 0$*,* $\beta_q(n)$ *is a decreasing function of* n*.*

4 The Exponential Efron-Stein Inequality

The main tool for our work is an extension of the celebrated Efron-Stein inequality (Efron and Stein 1981; Steele 1986),

[2] This should not be mistaken to taking a supremum over any subset of the dropped quantities.

to a stronger version known as the exponential Efron-Stein inequality (Boucheron, Lugosi, and Massart 2003). We start by introducing the Efron-Stein inequality and some variations. Let $f : \mathcal{X}^n \longmapsto \mathbb{R}$ be a real-valued function of n variables, where $\mathcal{X}$ is a measurable space. Let $X_1, \dots, X_n$ be independent (not necessarily identically distributed) RVs taking values in $\mathcal{X}$ and define the RV $Z = f(X_1, \dots, X_n) \equiv f(\mathcal{S}_n)$. Define the shorthand for the conditional expectation $\mathbb{E}_{-i}Z \doteq \mathbb{E}\left[Z|\mathcal{S}_n^{-i}\right]$, where $\mathcal{S}_n^{-i}$ is defined as in the previous section. Informally, $\mathbb{E}_{-i}Z$ "integrates" Z over X_i and *also over any other source of randomness* in Z except $\mathcal{S}_n^{-i}$. For every $i = 1, \dots, n$, let X_i' be an independent copy from X_i, and let $Z_i' = f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$. The Efron-Stein inequality bounds the variance of Z as shown in the following theorem.

Theorem 1 (Efron-Stein Inequality – Replacement Case)**.** *Let $V = \sum_{i=1}^n (Z - \mathbb{E}_{-i}Z)^2$. Under the settings described in this section, it holds that*

$$\mathbb{V}[Z] \leq \mathbb{E}V = \tfrac{1}{2}\sum_{i=1}^{n} \mathbb{E}[(Z - Z_i')^2].$$

The proof of Theorem 1 can be found in (Boucheron, Lugosi, and Massart 2004). Another variant of the Efron-Stein inequality that is more useful for our context, is concerned with the removal of one example from $\mathcal{S}_n$. To state the result, let $f_i : \mathcal{X}^{n-1} \longmapsto \mathbb{R}$, for $1 \leq i \leq n$, be an arbitrary measurable function, and define the RV $Z_{-i} = f_i(\mathcal{S}_n^{-i})$. Then, the Efron-Stein inequality can be also stated in the following interesting form (Boucheron, Lugosi, and Massart 2004).

Corollary 1 (Efron-Stein Inequality – Removal Case)**.** *Assume that $\mathbb{E}_{-i}[Z_{-i}]$ exists for all $1 \leq i \leq n$, and let $V_{DEL} = \sum_{i=1}^n (Z - Z_{-i})^2$. Then it holds that*

$$\mathbb{V}[Z] \;\leq\; \mathbb{E}V \;\leq\; \mathbb{E}V_{DEL}\,. \tag{3}$$

4.1 An Exponential Efron-Stein Inequality

The work of Boucheron, Lugosi, and Massart (2003) has focused on controlling the tail of general functions of independent RVs in terms of the tail behavior of Efron-Stein variance-like terms such as V and V_{DEL}, as well as other terms known as V^+ and V^-. The variance-like terms V, V^+ and V^- measure the sensitivity of a function of n independent RVs w.r.t the *replacement* of one RV from the n independent RVs. The term V_{DEL} on the other hand, measures the sensitivity of a function of n independent RVs w.r.t the *removal* of one RV from the n independent RVs. In this work, we favor V_{DEL} over the other terms since it is more suitable for our choice of stability coefficient (the L_q stability), which is also a removal version. The removal version of stability is preferred as it is more natural in the learning context where one is given a fixed sample. In particular, the removal version seems to be a better fit when it comes to empirically estimating stability (which is an interesting future direction), where working with the replacement version will need extra data, or extra assumptions.

The tail of a RV is often controlled through bounding the logarithm of the moment generating function (MGF) of the RV. This is known as the *cumulant generating function* (CGF) of the RV and is defined as

$$\psi_Z(\lambda) \doteq \log \mathbb{E}\left[\exp(\lambda Z)\right], \tag{4}$$

where $\lambda \in \operatorname{dom}(\psi_Z) \subset \mathbb{R}$ and belongs to a suitable neighborhood of zero. The main result of (Boucheron, Lugosi, and Massart 2003) bounds ψ_Z in terms of the MGF for V, V^+ and V^-, but not in terms of the MGF for V_{DEL}. Since we are particularly interested in the RV V_{DEL}, the following theorem bounds the tail of ψ_Z in terms of the MGF for V_{DEL}. The proof is given in the Appendix.

Theorem 2. *Let Z, V_{DEL} be defined as in Corollary 1 and assume that $|Z - Z_{-i}| \leq 1$ almost surely for all i. For all $\theta > 0$, such that $\lambda \in (0, 1]$, $\theta\lambda < 1$, and $\mathbb{E}e^{\lambda V_{DEL}} < \infty$, the following holds*

$$\log \mathbb{E}\left[\exp\left(\lambda(Z - \mathbb{E}Z)\right)\right] \leq \tfrac{\lambda\theta}{(1-\lambda\theta)} \log \mathbb{E}\left[\exp\left(\tfrac{\lambda V_{DEL}}{\theta}\right)\right]. \tag{5}$$

Theorem 2 states that the CGF of the centered RV $Z - \mathbb{E}Z$ is upper bounded by the CGF of the RV V_{DEL}. Hence, when V_{DEL} behaves "nicely", the tail of Z can be controlled. The value of θ in the upper bound is a free parameter that can be optimized to give the tightest bound.

For Theorem 2 to be useful in our context, further control is required to upper bound the tail of the RV V_{DEL}. Our approach to control the tail of V_{DEL} will be, again, through its CGF. In particular, we aim to show that when V_{DEL} is a sub-gamma RV (defined shortly) we can obtain a high probability tail bound on the deviation of the RV Z. The obtained tail bound will be instrumental in deriving the exponential tail bound for the deleted estimate.

4.2 Sub-Gamma Random Variables

We follow here the notation of (Boucheron, Lugosi, and Massart 2013). A real valued centered RV X is said to be *sub-gamma* on the right tail with variance factor v and scale parameter c if for every λ such that $0 < \lambda < 1/c$, the following holds

$$\psi_X(\lambda) \;\;\leq\;\; \frac{\lambda^2 v}{2(1 - c\lambda)}\,. \tag{6}$$

This is denoted by $X \in \Gamma_+(v, c)$. Similarly, X is said to be a sub-gamma RV on the left tail with variance factor v and scale parameter c if $-X \in \Gamma_+(v, c)$. This is denoted as $X \in \Gamma_-(v, c)$. Finally, X is simply a sub-gamma RV with variance factor v and scale parameter c if both $X \in \Gamma_+(v, c)$ and $X \in \Gamma_-(v, c)$. This is denoted by $X \in \Gamma(v, c)$.

The sub-gamma property can be characterized in terms of moments conditions or tail bounds. In particular, if a centered RV $X \in \Gamma(v, c)$, then for every $t > 0$,

$$\mathbb{P}\left[X > \sqrt{2vt} + ct\right] \vee \mathbb{P}\left[-X > \sqrt{2vt} + ct\right] \leq e^{-t}\,, \tag{7}$$

where $a \vee b = \max(a, b)$. The following theorem from (Boucheron, Lugosi, and Massart 2013) characterizes this notion more precisely.

Theorem 3. *Let X be a centered RV. If for some $v > 0$, $c \geq 0$, and for every $t > 0$,*

$$\mathbb{P}\left[X > \sqrt{2vt} + ct\right] \vee \mathbb{P}\left[-X > \sqrt{2vt} + ct\right] \leq e^{-t} \ , \quad (8)$$

then for every integer $q \geq 1$

$$\begin{aligned} \|X\|_{2q} &\leq \ (q!A^q + (2q)!B^{2q})^{1/2q} \\ &\leq \ \sqrt{16.8qv} \vee 9.6qc \ \ \leq \ 10(\sqrt{qv} \vee qc) \ , \end{aligned}$$

where $A = 8v$, $B = 4c$. Conversely, if for some positive constants u and w, for any integer $q \geq 1$,

$$\|X\|_{2q} \ \leq \ \sqrt{qu} \vee qw \ ,$$

then $X \in \Gamma(v, c)$ with $v = 4(1.1u + 0.53w^2)$ and $c = 1.46w$, and therefore (8) *also holds.*

The reader may notice that Theorem 3 is slightly different than the version in the book of Boucheron, Lugosi, and Massart (2013). Our extension is based on simple (and standard) calculations that are merely for convenience with respect to our purpose.

4.3 An Exponential Tail Bound for Z

In this section we assume that $V_{\text{DEL}} - \mathbb{E}V_{\text{DEL}}$ is a sub-gamma RV with variance factor $v > 0$, scale parameter $c \geq 0$, $\lambda > 0$, and $c\lambda < 1$. Hence, from inequality (6) it holds that

$$\begin{aligned} \psi_{V_{\text{DEL}} - \mathbb{E}V_{\text{DEL}}}(\lambda) &= \log \mathbb{E}\left[\exp(\lambda(V_{\text{DEL}} - \mathbb{E}V_{\text{DEL}}))\right] \\ &\leq \tfrac{1}{2}\lambda^2 v (1 - c\lambda)^{-1} \ . \end{aligned}$$

The sub-gamma property of V_{DEL} provides the desired control on its tail. That is, after arranging the terms of the above inequality, the CGF of V_{DEL} which controls the tail of V_{DEL}, is upper bounded by the deterministic quantities: $\mathbb{E}V_{\text{DEL}}$, the variance v, and the scale parameter c.

It is possible now to use the sub-gamma property of V_{DEL} in the result of the exponential Efron-Stein inequality in Theorem 2. In particular, the following lemma gives an exponential tail bound on the deviation of a function of independent RVs, i.e. $Z = f(X_1, \ldots, X_n)$, in terms of $\mathbb{E}V_{\text{DEL}}$, the variance factor v, and the scale parameter c. This lemma will be our main tool to derive the exponential tail bound on the DEL estimate. The proof is given in the Appendix.

Lemma 1. *Let Z, Z_{-i}, V_{DEL} be as in Corollary 1. If $V_{DEL} - \mathbb{E}V_{DEL}$ is a sub-gamma RV with variance parameter $v > 0$ and scale parameter $c \geq 0$, then for any $\delta \in (0, 1)$, $a > 0$, with probability $1 - \delta$,*

$$|Z - \mathbb{E}Z| \leq \tfrac{2}{3}(ac + 1/a)\log\left(\tfrac{1}{\delta}\right) + 2\sqrt{(\mathbb{E}V_{DEL} + a^2 v/2)\log\left(\tfrac{1}{\delta}\right)} .$$

Parameter a is a free parameter that can be optimized to give the tightest possible bound. In particular, a can be chosen to provide the appropriate scaling for the RV Z such that the bound goes to zero as fast as possible. A typical choice of a would be the inverse standard deviation of Z.

5 Main Result

In this section we derive the main result of this paper; an exponential tail bound for the concentration of the estimated risk, developed in terms of the deleted estimate, using the weak, distribution dependent and computationally tractable notion of L_q stability from Definition 2, and the Exponential Efron-Stein inequality from Lemma 1. In particular, we are interested in the concentration of the following RV

$$|\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \ .$$

To bound this RV, we decompose into three terms

$$\left|\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})\right| \ \leq \ \mathrm{I} + \mathrm{II} + \mathrm{III} \ , \qquad (9)$$

where

$$\begin{aligned} \mathrm{I} &= |\mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - \widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)| \ , \\ \mathrm{II} &= |\mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \ , \quad \text{and} \\ \mathrm{III} &= |\mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)| \ . \end{aligned}$$

If the three terms in the RHS of (9) are properly upper bounded, we will have the desired final high probability bound. Terms I and II shall be bounded using the exponential Efron-Stein inequality in Lemma 1. Further, we hope that the final upper bounds can be in terms of the L_q stability coefficient of A. Term III, however, is non-random and shall be directly bounded using some L_q stability coefficient.

5.1 Upper Bounding Term I

We begin by deriving an upper bound for term I in the RHS of (9). This is the deviation $|\mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - \widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)|$. Note that $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$ is a function of n independent random variables, and hence the Exponential Efron-Stein inequality from Lemma 1 seems to be applicable to bound this deviation. Following our two-steps plan to use Lemma 1, we define the random variables Z and Z_{-i} as follows

$$\begin{aligned} Z &= \widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right) \\ Z_{-i} &= \frac{1}{n-1}\sum_{\substack{j=1 \\ j \neq i}}^{n} \ell\left(\mathtt{A}(\mathcal{S}_n^{-i,-j}), X_j\right) , \end{aligned} \qquad (10)$$

where $\mathcal{S}_n^{-i,-j}$ indicates the removal of examples X_i and X_j from $\mathcal{S}_n$. Recall that $V_{\text{DEL}} = \sum_i (Z - Z_{-i})^2$, and given the definition of Z and Z_{-i} in (10), we need to show that V_{DEL} is a sub-gamma RV and derive a bound on $\mathbb{E}V_{\text{DEL}}$. This can be done by bounding the higher order moments of V_{DEL} as stated in the following lemma. The proof is given in the Appendix.

Lemma 2. *Let Z, Z_{-i} be defined as in* (10)*, and let $V_{DEL} = \sum_{i=1}^{n}(Z - Z_{-i})^2$. Then for any real $q \geq 1/2$ and $n \geq 2$, the following holds*

$$\|V_{DEL}\|_{2q} \ \leq \ n\beta_{4q}^2(n-1) \ , \qquad (11)$$

and hence

$$\mathbb{E}V_{DEL} \ \leq \ n\beta_2^2(n-1) \ .$$

Lemma 2 gives the desired upper bound for the higher order moments of V_{DEL} including the upper bound for $\mathbb{E}V_{\text{DEL}}$. To use Lemma 1, it remains to show that V_{DEL} is a sub-gamma RV according to the characterization in Theorem 3. However, since our results are for a *generic* learning rule $\mathtt{A}$ with minimal knowledge on its stability, we need to postulate the following mild assumption on the behavior of $(n\beta_{4q}^2(n-1))_{q\geq 1}$. Once $\mathtt{A}$ is specified, this assumption will not be needed since an upper bound can be realized for β_{4q}^2. For instance, as shown in (Celisse and Guedj 2016), and for the bounded ridge regression case, β_q is upper bounded by the q-norm of the response variable Y.

Assumption 2. $\exists\, u_1, w_1 \geq 0$ *s.t. for any integer* $q \geq 1$*, it holds that* $n\beta_{4q}^2(n-1) \leq \sqrt{qu_1} \vee qw_1$.

From Assumption 2, it follows immediately that V_{DEL} is a sub-gamma RV as stated in the following corollary.

Corollary 2. *Using the previous definitions, and under Assumption 2,* $V_{\mathit{DEL}} \in \Gamma(v_1, c_1)$*, where* $v_1 = 4(1.1u_1 + 0.53w_1^2)$ *and* $c_1 = 1.46w_1$.

The statement of Corollary 2 follows from Lemma 2, and using Assumption 2 and Theorem 3. Plugging the result of Lemma 2 and Corollary 2 into Lemma 1 gives the desired final upper bound for Term I in the RHS of (9).

Lemma 3. *Suppose that Assumption 2 holds and* $n \geq 2$*. Then for any* $\delta \in (0,1)$ *and* $a > 0$*, with probability* $1-\delta$ *the following holds*

$$\begin{aligned}\tilde{R}_I &= |\mathbb{E}\widehat{R}_{\mathit{DEL}}(\mathtt{A}, \mathcal{S}_n) - \widehat{R}_{\mathit{DEL}}(\mathtt{A}, \mathcal{S}_n)| \\ &\leq \tfrac{2}{3}(1.46aw_1 + \tfrac{1}{a})\log\left(\tfrac{1}{\delta}\right) \\ &\quad + 2\sqrt{(n\beta_2^2(n-1) + \rho_1(u_1, w_1))\log\left(\tfrac{1}{\delta}\right)}\ ,\end{aligned}$$

where $\rho_1(u_1, w_1) = 2.2a^2u_1 + 1.07a^2w_1^2$.

Consider now the choice of a in the context of how it may scale with n and its impact on the behavior of this bound. First, note that u_1 and w_1 are controlled by $n\beta_{4q}^2(n-1)$, and from Assumption 1, we assume that $\beta_{4q}^2(n-1)$ is a decreasing function of n. If, for example, $n\beta_2^2(n-1) \sim \frac{1}{n^p}$ for some $p > 0$, then $u_1 \sim n^{-2p}$, $w_1 \sim n^{-p}$, and $w_1 \approx \sqrt{u_1}$. The terms in the bound that depend on a scale as $\frac{a}{n^p} + \frac{1}{a}$ with n. Hence, choosing $a = n^{p/2}$, or $a = w_1^{-1/2}$, makes both, the a dependent term, as well as the whole bound, scale with $n^{-p/2}$ as a function of n; i.e. the bound scales as $w_1^{1/2}$, and $w_1^{1/2} = o(1)$ as $n \to \infty$. This translates to $n\beta_{4q}^2(n-1) = o(1)$ as $n \to \infty$; i.e. $\beta_2(n-1) = o(n^{-1/2})$ which is sufficient for the consistency of $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$. A similar condition for consistency was also identified by Bousquet and Elisseeff (2002) and Celisse and Guedj (2016).

5.2 Upper Bounding Term II

Consider now term II in inequality (9). This is the deviation $|\mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) - R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)|$. Note that $R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)$ is a function of n independent RVs, and therefore, Lemma 1 will be our tool to bound this deviation. Following the steps for upper bounding Term I in the previous section, we need to define the RVs Z and Z_{-i}, and show that V_{DEL} is a sub-gamma RV. Let the RVs Z and Z_{-i} be defined as follows

$$\begin{aligned} Z &= R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) \\ Z_{-i} &= R\left(\mathtt{A}(\mathcal{S}_n^{-i}), \mathscr{P}\right) . \end{aligned} \tag{12}$$

Similar to Lemma (2) we have the following result:

Lemma 4. *Let* Z *and* Z_{-i} *be defined as in* (12) *and let* $V_{\mathit{DEL}} = \sum_{i=1}^n (Z - Z_{-i})^2$*. Then for any real* $q \geq 1/2$*, and* $n \geq 2$*, the following holds*

$$\|V_{\mathit{DEL}}\|_{2q} \ \leq\ n\beta_{4q}^2(n,1)\ , \tag{13}$$

and hence

$$\mathbb{E}V_{\mathit{DEL}} \ \leq\ n\beta_2^2(n,1)\ . \tag{14}$$

For the same reason we made Assumption 2, we need to make the following assumption.

Assumption 3. $\exists\, u_2, w_2 \geq 0$ *s.t. for any integer* $q \geq 1$*, it holds that* $n\beta_{4q}^2(n,1) \leq \sqrt{qu_2} \vee qw_2$.

Corollary 3. *Using the previous definitions, and under Assumption 3,* $V_{\mathit{DEL}} \in \Gamma(v_2, c_2)$*, where* $v_2 = 4(1.1u_2 + 0.53w_2^2)$ *and* $c_2 = 1.46w_2$.

The steps to derive the final bound for Term II are exactly the same derivation steps for the previous bound. The final bound is given by the following lemma which plugs in the results of Lemma (4) and Corollary (3) into Lemma 1.

Lemma 5. *Suppose that Assumption 3 holds and* $n \geq 2$*. Then for any* $\delta \in (0,1)$ *and* $a > 0$*, with probability* $1-\delta$ *the following holds*

$$\begin{aligned}\tilde{R}_{\mathit{II}} &= |\mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) - R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)| \\ &\leq \tfrac{2}{3}(1.46aw_2 + \tfrac{1}{a})\log\left(\tfrac{1}{\delta}\right) \\ &\quad + 2\sqrt{(n\beta_2^2(n) + \rho_2(u_2, w_2))\log\left(\tfrac{1}{\delta}\right)}\ ,\end{aligned}$$

where $\rho_2(u_2, w_2) = 2.2a^2u_2 + 1.07a^2w_2^2$.

Concerning the choice of a, the discussion after Lemma 3 applies.

5.3 Upper Bounding Term III

For term III in inequality (9) there are no random quantities to account for since both terms in the modulus are expectations of RVs. Hence, an upper bound on this deviation will always hold.

Lemma 6. *Using the previous setup and definitions, let* $\mathtt{A}$ *be a learning rule with* L_2 *stability coefficient* $\beta_2(n)$*. Then for* $n \geq 2$*, the following holds*

$$|\mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}\widehat{R}_{\mathit{DEL}}(\mathtt{A}, \mathcal{S}_n)| \ \leq\ \beta_2(n)\ .$$

5.4 Main Result

We arrive now to the main result of this work, namely an exponential tail bound for the concentration of the estimated risk, expressed in terms of the deleted estimate, for a general learning rule using the notion of L_q stability.

Theorem 4. *Let $\mathcal{X}$, $\mathcal{H}$ and ℓ be as previously defined. Let $\mathcal{S}_n$ be the dataset defined in Section 2.1, where $n \geq 2$. Let $\widehat{R}_{DEL}\,(\mathtt{A}, \mathcal{S}_n)$ be the deleted estimate defined in Eq.* (2)*, and $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$ be the risk for hypothesis $\mathtt{A}(\mathcal{S}_n)$. Then, under Assumption 2 and Assumption 3, for $\delta \in (0, 1)$ and $a > 0$, with probability $1 - \delta$ the following holds*

$$\begin{aligned}\tilde{R}_{DEL} &= |\widehat{R}_{DEL}\,(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \\ &\leq \beta_2(n) + 4\sqrt{(n\beta_2^2(n-1) + C_1)\log\left(\tfrac{1}{\delta}\right)} + C_2 \log\left(\tfrac{1}{\delta}\right),\end{aligned}$$

where $C_1 = 2.2a^2u_1 + 1.07a^2w_1^2$, and $C_2 = \frac{4}{3}(1.46aw_1 + \frac{1}{a})$.

The proof of Theorem 4 starts by plugging the results of Lemma 3, Lemma 5, and Lemma 6 into inequality (9). Next, to simplify the expression and improve the presentation of the final result, we proceed as follows. From Assumption 1 we have that $\beta_2^2(n-1) \geq \beta_2^2(n)$. Combining this with Assumption 2 and Assumption 3, we expect that $(\sqrt{qu_1} \vee qw_1) \geq (\sqrt{qu_2} \vee qw_2)$, and hence $\rho_1(u_1, w_1) \geq \rho_2(u_2, w_2)$. This implies that the RHS for the inequality in Lemma 3 upper bounds the RHS for the inequality in Lemma 5. Thus, replacing $\beta_2^2(n)$ with $\beta_2^2(n-1)$, $\rho_2(u_2, w_2)$ with $\rho_1(u_1, w_1)$, $\frac{2}{3}(1.46aw_2 + \frac{1}{a})$ with $\frac{2}{3}(1.46aw_1 + \frac{1}{a})$, and summing all the terms yields the final bound in Theorem 4. Concerning the choice of a, the discussions that follow Lemma 3 and Lemma 5 apply here.

Discussion: Consider the three terms that constitute the bound in Theorem 4 and note that all the terms depend on the stability of the learning rule. While the first term is obvious in this regard, the second term has an explicit dependence on the stability through $\beta_2^2(n-1)$, as well as an implicit dependence through the constant C_1 which itself is dependent on the higher order moments of the L_q stability RV β_{4q}^2. Recall from Assumption 2 that u_1 and w_1 are dependent on β_{4q}^2. The same applies for the third term where constant C_2 also depends on w_1. Thus, as the stability is improving (i.e. smaller β_{4q}^2), w_1 and u_1 become smaller, and the whole bound becomes tighter. Note that from C_2, there is a small factor of $\frac{4}{3a}\log\left(\frac{1}{\delta}\right)$ that cannot be avoided even for very stable learning rules.

At a higher level, the proof technique followed from Lemma 1 to the final bound in Theorem 4 can be summarized as follows: in order to control the concentration of the random quantity $\widehat{R}_{\text{DEL}}\,(\mathtt{A}, \mathcal{S}_n)$ around the true risk, one has to control the tails (or the higher order moments) of $\widehat{R}_{\text{DEL}}\,(\mathtt{A}, \mathcal{S}_n)$. In turn, to control the tails of $\widehat{R}_{\text{DEL}}\,(\mathtt{A}, \mathcal{S}_n)$, one has to control the tails (or the higher order moments) of another random variable, V_{DEL}, which turns to be the L_q stability coefficients of the learning rule.

Before closing, we believe it can be useful to qualitatively compare our bound in Theorem 4 with the exponential bound for the deleted estimate obtained by Bousquet and Elisseeff (2002) (Theorem 12) for the regression case. To make the comparison easier, we first state their result using our notation.

Theorem 5. *Let $\mathtt{A}$ be a learning rule with uniform stability β_u (see Section 3) with respect to the loss function ℓ such that $\forall X \sim \mathscr{P}$, and $\forall \mathcal{S} \sim \mathscr{P}^n$, it holds that $0 \leq \ell\,(\mathtt{A}(\mathcal{S}_n), X) \leq M$. Then, for any $n \geq 1$, and any $\delta \in (0, 1)$, with probability $1 - \delta$, the following holds*

$$\begin{aligned}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \widehat{R}_{DEL}\,(\mathtt{A}, \mathcal{S}_n) \leq \beta_u(n) + 4n\beta_u(n)\sqrt{\frac{\log(1/\delta)}{2n}} \\ + M\sqrt{\frac{\log(1/\delta)}{2n}}\ .\end{aligned}$$

The bound in Theorem 5 has three main terms; the first two terms are dependent on the uniform stability of the learning rule, and a third term that only depends on the loss function ℓ and the sample size n. When β_u scales as $1/n$ the bound becomes tight, however, even for very stable learning rules, the third term cannot be avoided. The first term in the RHS of Theorem 5 corresponds to the first term in our bound in Theorem 4 where both terms are derived from the same quantity, i.e. $|\mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)|$, but under different notions of stability. The second term the in RHS of Theorem 5, which can be written as $4\sqrt{n^2\beta_u^2\log(1/\delta)/2n}$ resembles our second term $4\sqrt{n\beta_2^2(n-1)\log(1/\delta)}$ albeit without the additional term of $C_1\log(1/\delta)$. Indeed, for a learning rule that satisfies the strong notion of uniform stability, this term will make the final bound tighter than our second term. However, relaxing the requirements of uniform stability by adopting the notion of L_q stability instead, the additional terms $\sqrt{C_1\log(1/\delta)}$ and $C_2\log(1/\delta)$ kick in our final bound. These terms are due to the higher order moments of the RV $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$ which translate to the higher order moments of the RV $|\ell\,(\mathtt{A}(\mathcal{S}_n), X) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X\right)|$; i.e. the L_q stability coefficients. In some sense, these additional terms due to the higher order moments of stability, seem to compensate for the gap between uniform stability and L_q stability to ensure the proper concentration of the estimated risk around its expectation. Last, the terms $M\sqrt{\log(1/\delta)/2n}$ in Theorem 5 and $\frac{4}{3a}\log(1/\delta)$ in Theorem 4, which cannot be avoided even for very stable learning rules, somehow correspond to the bias of the estimator but under two different notions of stability.

6 Conclusion

Our work here considers the gap between two regimes of stability-based generalization results; (*i*) exponential generalization bounds based on strong notions of stability which are distribution independent and computationally intractable, such as uniform stability, and (*ii*) polynomial generalization bounds based on weaker notions of stability but are distribution dependent and computationally tractable such as hypothesis stability and L_q stability. Using the exponential Efron-Stein inequality we were able to bridge this gap by deriving an exponential concentration bound for L_q stable learning rules, where the loss of the learning rules is expressed in terms of the deleted estimate. We believe that our result is one step forward on two fronts; (*i*) computing empirical tight confidence intervals for the expected loss of a learning rule where the confidence interval holds with high probability; and (*ii*) understanding the role of stability in the concentration of different empirical loss estimates around their expectations.

References

Agarwal, S., and Niyogi, P. 2009. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research (JMLR)* 10:441–474.

Audibert, J.-Y.; Munos, R.; and Szepesvári, C. 2007. Tuning bandit algorithms in stochastic environments. In Hutter, M.; Servedio, R. A.; and Takimoto, E., eds., *Algorithmic Learning Theory*, 150–165. Springer Berlin Heidelberg.

Bassily, R.; Nissim, K.; Smith, A.; Steinke, T.; Stemmer, U.; and Ullman, J. 2016. Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC'16, 1046–1059. ACM.

Boucheron, S.; Lugosi, G.; and Massart, P. 2003. Concentration inequalities using the entropy method. *The Annals of Probability* 31(3):1583–1614.

Boucheron, S.; Lugosi, G.; and Massart, P. 2004. Concentration inequalities. In Bousquet, O.; von Luxburg, U.; and Rätsch, G., eds., *Advanced Lectures in Machine Learning*. Springer. 208–240.

Boucheron, S.; Lugosi, G.; and Massart, P. 2013. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

Bousquet, O., and Elisseeff, A. 2002. Stability and generalization. *Journal of Machine Learning Research (JMLR)* 2:499–526.

Celisse, A., and Guedj, B. 2016. Stability revisited: new generalisation bounds for the leave-one-out. *ArXiv e-prints* (1608.06412).

Devroye, L., and Wagner, T. 1979a. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. on Information Theory* 25(2):202–207.

Devroye, L., and Wagner, T. 1979b. Distribution-free performance bounds for potential function rules. *IEEE Trans. on Information Theory* 25(5):601–604.

Devroye, L.; Györfi, L.; and Lugosi, G. 1996. *A Probabilistic Theory of Pattern Recognition*. Springer.

Efron, B., and Stein, C. M. 1981. The jackknife estimate of variance. *The Annals of Statistics* 9(3):586–596.

Elisseeff, A.; Evgeniou, T.; and Pontil, M. 2005. Stability of randomized learning algorithms. *Journal of Machine Learning Research (JMLR)* 6:55–79.

Hardt, M.; Recht, B.; and Singer, Y. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML'16, 1225–1234. JMLR.org.

Holden, S. B. 1996. Pac-like upper bounds for the sample complexity of leave-one-out cross-validation. In *Proc. of the Ninth Annual Conference on Computational Learning Theory*, COLT '96, 41–50. ACM.

Kale, S.; Kumar, R.; and Vassilvitskii, S. 2011. Cross validation and mean-square stability. In *In Proceedings of the Second Symposium on Innovations in Computer Science ICS'2011*.

Kearns, M., and Ron, D. 1999. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation* 11(6):1427–1453.

Kutin, S., and Niyogi, P. 2002. Almost-everywhere algorithmic stability and generalization error. In *Proc. of Uncertainty in Artificial Intelligence (UAI)*, 275–282.

London, B.; Huang, B.; and Getoor, L. 2016. Stability and generalization in structured prediction. *Journal of Machine Learning Research (JMLR)* 17(222):1–52.

Lugosi, G., and Pawlak, M. 1994. On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Trans. on Information Theory* 40(2):475–481.

Maurer, A. 2005. Algorithmic stability and meta-learning. *Journal of Machine Learning Research (JMLR)* 6:967–994.

Mnih, V.; Szepesvári, C.; and Audibert, J.-Y. 2008. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, ICML'08, 672–679. ACM.

Mukherjee, S.; Niyogi, P.; Poggio, T.; and Rifkin, R. 2006. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics* 25:161–193.

Rakhlin, A.; Mukherjee, S.; and Poggio, T. 2005. Stability results in learning theory. *Analysis and Applications* 03(04):397–417.

Rogers, W. H., and Wagner, T. J. 1978. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics* 6(3):506–514.

Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; and Sridharan, K. 2010. Learnability, stability and uniform convergence. *Journal of Machine Learning Research (JMLR)* 11:2635–2670.

Steele, J. M. 1986. An Efron-Stein inequality for nonsymmetric statistics. *The Annals of Statistics* 14(2):753–758.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.

Xu, H.; Caramanis, C.; and Mannor, S. 2012. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 34(1):187–193.

Zhang, Y. 2015. Multi-task learning and algorithmic stability. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 3181–3187. AAAI Press.